통계적 방법을 이용한 웜 전파 모델링

정회원 우경문*, 종신회원 김종권**

Internet Worm Propagation Modeling using a Statistical Method

Kyung-moon Woo* *Regular Member*, Chong-kwon Kim** *Lifelong Member*

요 약

인터넷 웜은 컴퓨터 네트워크를 이용하여 자기 자신을 자동으로 복제해서 전파하는 프로그램이다. 컴퓨터간의 네트워크 연결이 증가함에 따라 인터넷 웜은 급격해 확산되었고 큰 위협으로 남아있다. 코드 레드, 님다, 슬레머 같은 인터넷 웜의 특성과 이들의 활동을 억제하는 방법을 찾기 위해서 웜이 전파되는 특성을 연구하려는 많은 시도가 있었다. 네트워크 특징들이 인터넷 웜 전파에 미치는 영향은 모델의 간단성과 유사성 때문에 주로 의학계에서 사용되는 전염병 전파 모델을 이용하여 모델링이 되었다. 이런 의학계 모델링은 널리 사용되면서 여러 개선된 모델들이 다양하게 제안되었다. 우리는 이전의 제안된 모델들의 문제점을 분석한 후 통계적 방법을 사용하여 정확도를 높이는 새로운 방법의 웜 전파 모델링을 제안한다.

Key Words : Internet Worm, Propagation modeling, Occupancy Problem

ABSTRACT

An Internet worm is a self-replicating malware program which uses a computer network. As the network connectivity among computers increases, Internet worms have become widespread and are still big threats. There are many approaches to model the propagation of Internet worms such as Code Red, Nimda, and Slammer to get the insight of their behaviors and to devise possible defense methods to suppress worms' propagation activities. The influence of the network characteristics on the worm propagation has usually been modeled by medical epidemic model, named SI model, due to its simplicity and the similarity of propagation patterns. So far, SI model is still dominant and new variations of the SI model, called SI-style models, are being proposed for the modeling of new Internet worms. In this paper, we elaborate the problems of SI-style models and then propose a new accurate stochastic model using an occupancy problem.

Ⅰ. Introduction

As the network connectivity among computers increases, Internet worms have become widespread and are still big threats. Conficker, also called Downup or Kido, is one of the most important worm outbreaks and became known to the public in November 2008[1]. Conficker worm attacks the vulnerability of the Microsoft windows family. Conficker attacks TCP 445 port where the vulnerable Microsoft's software runs and has evolved from Conficker A to E[1]. According to

※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0027410).

* 서울대학교 컴퓨터공학부 정보통신 연구실(kwoo@popeye.snu.ac.kr), ** 서울대학교 컴퓨터공학부 교수(ckim@snu.ac.kr)

논문번호 : KICS2011-12-635, 접수일자 : 2011년 12월 28일, 최종논문접수일자 : 2012년 3월 9일

SRI's recent technical report, infection size of Conficker A and B are 1 million and 3 million hosts each[2].

Most worms use random scanning to select victims. There are many approaches to model this random scanning and worm propagation so far. Medical epidemic model, called SI model, is applied in worm propagation for the first time[3]. SI-style models are still dominant, and new variations are being proposed. However, by conducting comprehensive analysis and simulations, we found that these SI-Style models have fundamental weaknesses for modeling worm propagations accurately. To compensate the weaknesses, we propose a new stochastic model. By applying occupancy problem, we can model worm's random scanning behaviors in high accuracy.

In this paper, we focus on the estimation of the total number of infected nodes at peak while it is in an early propagation stage. Firstly, we show the problems of the already proposed SI-style models, originally developed for the medical analysis. To resolve these problems, we

Table 1. SI style models

Scheme	SIS with local area control[7]	SIRD model[8]	SIS-BD model[9]	P2P virus model[10]	SIHR model[11]	SAIC model[12]	SAIR model[13]
Title	Susceptible-Infected-Susceptible Virus Spread Model in 2-Dimension Regular Network under Local Area Control	General worm propagation model for wireless ad hoc networks	The SIS-BD model of computer virus spreading on internet	A Computer Virus Propagation Model in P2P Networks	The Worm Propagation Model and Control Strategy Based on Distributed Honeynet	Dynamical models for computer viruses	A modified epidemiological model for computer viruses
Publica-tion	ICNDS, 2009	ICCSIT, 2009	WiCom, 2007	IWETCS, 2009	IEEE ICCSSE, 2008	Computers & Security, 2008	Mathematical Problems in Engineering, 2009
Method	SIS model	SIR model	SIS model	SIS, SIR model	Two-factor model	SI model	SIR model
Domain	Internet Virus	Ad Hoc Network Worm	Internet Virus	P2P Virus	Internet Worm	Internet Virus	Internet virus
Pros.	Consider Local Area control	Modeling with markov chain Analysis of steady-state of mean field approximation	Enhance SIS model with birth & death	Fine grained model for p2p	Modeling with distributed honeynet	Update of SAI model with contaminated state	Update of SIR model with antidotal state
Cons.	Minor update from classical SIS model	No comparison with real data	Minor model update from classical SI model	Minor update from classical SIS model	Minor update from two-factor model	Minor model update from classical SI model	Minor model update from classical SIR model

propose a new model using occupancy problem which stochastically calculates all the possible scan activities of a worm. Our new stochastic model requires considerable computation overhead. Thus, we simplify this stochastic model into a simple model which dramatically reduces the computation overhead, yet shows similar accuracy. The rest of the paper is organized as follows: In Section 2, we review the related work on worm propagation. In Section 3, we show the accuracy problem of previous models, more specifically, SI-style models. In Section 4 we propose a new stochastic model and conduct model validation with a simulation program. In Section 5, we conclude our research and provide some discussion and directions for future work.

Ⅱ. Related Work

Originated in medical modeling, classical epidemic model was first appeared in the worm modeling in 2002[4]. In this model, every host is either Susceptible or Infected. This is why this model is called SI model. There are many variations of this model such as SIR[5], SAIR model[5], DSIR model[6], ADSIR model[5]. SI-style models are still widely being used to model worm or virus propagations. Table 1 shows various schemes which use SI-style models.

Table 2 shows other schemes which use non-SI models. Sellke[14] propose Brach process model, a Markov process that models population increase as a tree branches. This model provides

Table 2. NON-SI style models

Scheme	Branching process model[14]	N-intertwined Markov chain model[15]	Correlation Method[16]	LMF model[17]	Graph Theory model in Cellular[18]
Title	Modeling and Automated Containment of Worms	Virus Spread in Networks	Dynamical models for computer viruses Propagation	A Local Mean Field Analysis of Security Investments in Networks	A Social Network Based Patching Scheme for Worm Containment in Cellular Networks
Publication	Transactions on Dependable and Secure Computing, 2008	ToN, 2009	Mathematical Problems in Engineering, 2008	NetEcon, 2008	INFOCOM, 2009
Method	Branching process (Borel-Tanner distribution)	Mean field approximation	Linear autoregressive, Fourier series	Epidemic propagation + economic model	Graph Theory
Domain	Internet Worm	Internet Virus	Internet Virus	Internet Virus and Worm	Cellular Worm
Pros.	Propose worm containment system Consider Random scanning and local preference scanning	Modeling with markov chain Analysis of steady-state of mean field approximation	Topological factors can be considered	Adopt economic model for agents to captures network effects	Consideration of MMS Worm Patching model with graph theory
Cons.	No consideration of duplicated scans and concurrent scans	No comparison with real data	Inaccurate method	No comparison with real data	No detail consideration of worm propagation

a precise bound on the total number of scans that ensures that the worm will be finally diminished and calculates the probability that the total number of hosts that the worm infects is below a certain level. This model is developed for uniform scanning worms and extended to preference scanning worms. The authors also suggest the automatic worm containment strategy that prevents the propagation of a worm at its early stage. The problem of this approach is infection probability should be constant, but as infected nodes are increased, infection probability should be decreased because susceptive nodes are reduced.

Mieghem[15] proposes N-intertwined Markov chain model using mean field approximation, which models the virus spread in a bi-directional network specified by a symmetric adjacency matrix. Several upper bounds for the steady-state infection probabilities are presented. The exact Markov chain provides insight into the virus spread process, but this model is not suitable for the estimation of virus or worm propagation.

Ⅲ. Problems of Previous Work

Medical epidemic models have been used to model worm propagation in [7]-[13]. As mentioned in Section 1, traditional SI model has many problems which affect model accuracy: overlapping scan problem, concurrent scan problem, model mismatch problem when total population and vulnerable population are different.

3.1. Overlapping Scanning

Equation (1) shows the increase of infectious nodes in original SI model.

$$\frac{dI}{dt} = \beta IS, \beta = \frac{1}{L}, \qquad (1)$$

Where I is the number of infectious nodes, S is the number of susceptible nodes, and L is the total number of nodes on the Internet. If many infectious nodes try to scan, it is possible that a certain node is scanned by more than 2 at the same time. As the number of scanning node increases, the overlapping scan probability increases, which reduces the infection rate which is expected by SI model.

Fig. 1 shows the difference between the simulation and SI models. As the overlapping scans are not considered, SI model overestimated the infected nodes.

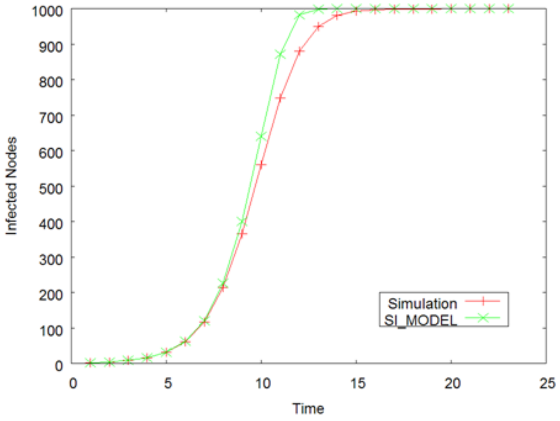

Fig. 1. Simulation and SI model comparison when total population (L) is 1000, and vulnerable population (N) is 1000.

3.2. Concurrent Scanning

Most worms use multi-thread to increase infection rates, but SI model cannot reflect concurrent scan activities. For example, suppose 2 cases: total population (L) is 1000, vulnerable population (N) is 1000, concurrent scan number is 1, and total population (L) is 10000, vulnerable population (N) is 1000, concurrent scan number is 10. In SI model, the results are the same.

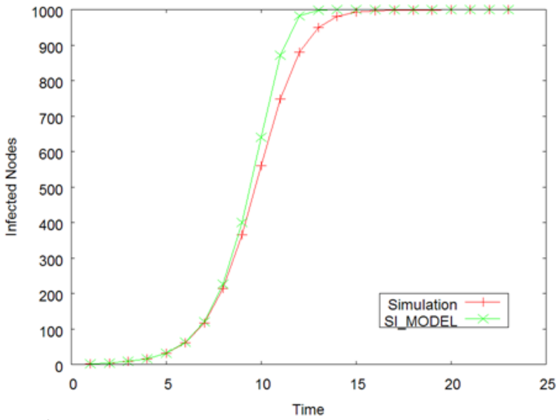

Fig. 2. Simulation and SI model comparison when total population (L) is 1000, vulnerable population (N) is 1000, and concurrent scan is 1.

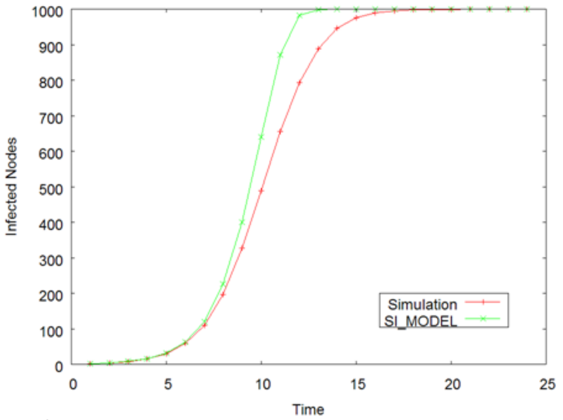

Fig. 3. Simulation and SI model comparison when total population (L) is 10000, vulnerable population (N) is 1000, and concurrent scan is 10.

Fig. 2 and 3 show the differences of infection curves when concurrent scan is applied in simulation, while SI models show the same infection curves.

Ⅳ. Model Proposal and Validation

We propose a new accurate stochastic model using occupancy problem. Occupancy problems deal with pairings of objects[22]. The basic occupancy problem is about placing m balls into n bins. We can consider a scan as a ball and each node as a bin.

Equation (2) and (3) show the number of infected nodes at time *t*, where $P_{t,i}$ is the probability that there are *j* infected nodes when time *t* and $P_{I=j}(X=i)$ is the probability that there are *i* newly infected nodes when current infected nodes are *j*.

Table 3. Notations

Symbol	Explanation
L	Total population (= in IPv4).
N	Total vulnerable population.
S	Susceptible population.
I	Initial infected nodes.
I(t)	Number of infected nodes at time t.
$P_{t,j}$	Probability that there are j infected nodes when time t.
$P_{I=j}(X=i)$	Probability that there are i newly infected nodes when current infected nodes are j.

$$I(t)=\sum_{i=I_0}^{\min(N,I_0\times 2^t)}\left[i\times\sum_{j=i_0}^{i}\{P_{t-1,j}\times P_{I=j}(X=i-j)\}\right] \tag{2}$$

$$P_{I=j}(X=i)={}_SC_i\sum_{k=i}^{j}\left[{}_jC_k\frac{(L-S)^{j-k}}{L^j}\left\{\sum_{r=0}^{k}(-1)^r\,{}_iC_r(i-r)^k\right\}\right] \tag{3}$$

Fig. 4 shows the accuracy of stochastic model. For the experiments, an Intel i5 (2.8Ghz) computer is used and worm propagation activities are simulated by random IP selection. We can see that our new stochastic model matches very well with simulation, while SI model shows over estimation.

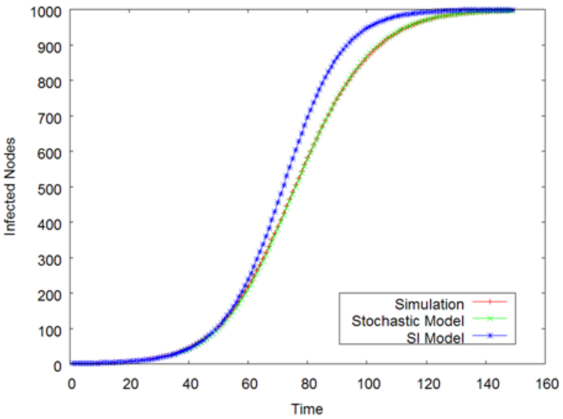

Fig. 4. Infected nodes when total population is 10000, and vulnerable population is 1000.

Ⅴ. Conclusion

SI-style models, originally developed for the medical analyses, are dominant models so far and new variations with additional considerations are being proposed in modeling worm or virus propagation. In this paper, we have showed the problems of already proposed SI-style models with 3 factors.

By applying occupancy problem, we have proposed a new stochastic model which shows very good accuracy. And we also have proposed a simple method which can reduce the computational overhead of the stochastic model, yet show similar accuracy.

References

[1] E. Aben, “Conficker/Conflicker/Downadup as seen from the UCSD Network Telescope,” http://www.caida.org/research/security/ms08-067/conficker.xml, 2009.

[2] Phillip Porras, Hassen Saıdi, and Vinod Yegneswaran, “A Foray into Conficker’s Logic and Rendezvous Points,” USENIX Workshop on Large-scale Exploits and Emergent Threats (LEET), 2009

[3] Kephart J O, White S R, “Directed-graph Epidemiological Models of Computer Viruses,” Proc. of the IEEE Computer Society Symposium on Research in Security and Privacy. Oakland, California, USA: IEEE Computer Society Press, 1991: 343-359.

[4] Stuart Staniford, Vern Paxson, Nicholas Weaver, “How to Own the internet in your spare time,” USENIX security, 2002

[5] Piqueira JRC, Navarro BF, Monteiro LHA, “Epidemiological models applied to viruses in computer networks,” journal of computer science, 2005.

[6] Dagon D, Zou CC, Lee W., “Modeling botnet propagation using time zones,” In: Proc. Of the 13thAnnualNetworkandDistributedSystemSecuritySymp.(NDSS2006).2006.

[7] ChangRui Guo, ShaoHong Cai, HaiPing Zhou, DaMin Zhang, “Susceptible-Infected-Susceptible Virus Spread Model in 2-Dimension Regular Network under Local Area Control,” International Conference on Networking and Digital Society, 2009.

[8] Chen Junhua, Wei Shengjun, Peng Wu, “General Worm Propagation Model for Wireless Ad Hoc Networks,” International Conference on Computer Science and Information Technology, 2009.

[9] Ben-hua Guo, Shao-hong Cai, “THE SIS-BD MODEL OF COMPUTER VIRUS SPREADING ON INTERNET,” Wireless Communications, Networking and Mobile Computing, 2007.

[10] Ming Liu, Lansheng Han*, Fan Hong, Mengsong Zou, “A Computer Virus Propagation Model in P2P Networks,” International Workshop on Education Technology and Computer Science, 2009.

[11] Narisa Zhao, Xianfeng Zhang, “The Worm Propagation Model and Control Strategy Based on Distributed Honeynet,” International Conference on Computer Science and Software Engineering, 2008.

[12] Jose R.C. Piqueira, Adolfo A. de Vasconcelos, Carlos E.C.J. Gabriel, Vanessa O. Araujo, “Dynamical models for computer viruses,” Computers & Security, 2008.

[13] Jose Roberto C. Piqueira, Vanessa O. Araujo, “A modified epidemiological model for computer viruses,” Applied Mathematics and Computation, 2009.

[14] Sarah H. Sellke, Ness B. Shroff, Saurabh Bagchi, “Modeling and Automated Containment of Worms,” Transactions on Dependable and Secure Computing, 2008.

[15] Piet Van Mieghem, Jasmina Omic, and Robert Kooij, “Virus Spread in Networks,” Transactions on Networking, 2009.

[16] Jose R. C. Piqueira and Felipe Barbosa Cesar, “Dynamical models for computer viruses Propagation,” Mathematical Problems in Engineering, 2008

[17] Marc Lelarge, Jean Bolot, “A Local Mean Field Analysis of Security Investments in Networks,” Proceedings of the 3rd international workshop on Economics of networked systems, 2008.

[18] Zhichao Zhu, Guohong Cao, Sencun Zhu, Supranamaya Ranjan and Antonio Nucci, “A Social Network Based Patching Scheme for Worm Containment in Cellular Networks,” INFOCOM, 2009.

우 경 문 (Kyung-moon Woo) 정회원

1995년 공군사관학교 전산과 (학사). 2007년 서울대학교 전기, 컴퓨터공학부(석사) 2012년 서울대학교 전기, 컴퓨터공학부(박사)

<관심분야> 무선랜, 이동통신, 네트워크 보안

김 종 권 (Chong-kwon Kim) 종신회원

1982년 미국 조지아 공과대학교 산업공학과(석사). 1987년 미국 일리노이 대학교 전산학과(박사).

1984년~1987년 IBM 산호세 연구소 연구조원. 1987년 1월~1991년 미국 Belcore 통신연구소 연구원. 1991년~현재 서울대학교 전기·컴퓨터공학부 교수.

<관심분야> 차세대인터넷, 초고속라우터, 이동통신